Multiple Objective Optimization in Reliability Demonstration Tests

LU LU and MINGYANG LI

University of South Florida, Tampa, FL 33620

CHRISTINE M. ANDERSON-COOK

Los Alamos National Laboratory, Los Alamos, NM 87545

Reliability demonstration tests are usually performed in product design or validation processes to demonstrate whether a product meets specified requirements on reliability. For binomial demonstration tests, the zero-failure test has been most commonly used due to its simplicity and use of minimum sample size to achieve an acceptable consumer's risk level. However, this test can often result in unacceptably high risk for producers as well as a low probability of passing the test even when the product has good reliability. This paper explicitly explores the interrelationship between multiple objectives that are commonly of interest when planning a demonstration test and proposes structured decision-making procedures using a Pareto front approach for selecting an optimal test plan based on simultaneously balancing multiple criteria. Different strategies are suggested for scenarios with different user priorities and graphical tools are developed to help quantify the trade-offs between choices and to facilitate informed decision making. Potential impacts of some subjective user inputs on the final decision are studied to offer insights and useful guidance for general applications.

Key Words: Bayesian Analysis; Consumer's Risk; Decision Making; Pareto Front; Producer's Risk; Trade-Offs.

1. Introduction

Reliability demonstration tests are commonly used in product development and validation processes to ensure a certain reliability requirement is met. For example, they can be used to demonstrate reliability performance of a new design or a modification of an existing design in the early stages of its product-development cycle. They can also be used to assess if reliability of an existing design required to operate in new environmental or operational conditions exceeds the minimum acceptable requirement prior to releasing it into service. The tests are usually performed at the system level and set up as pass/fail tests.

Demonstration tests have been broadly employed in many industries, including microelectronics, aerospace, and healthcare, to guide decisions of the acceptance of some products or designs. To implement the test, one needs to develop a test plan for demonstrating a certain reliability performance with some desired level of confidence based on the available budget and resources. This requires the practitioners to answer questions including how many units should be tested, for how long each unit needs to be tested and under what conditions, as well as the decision rule for a successful or failed test.

This paper considers the binomial demonstration test for nonrepairable systems, where a sample of units is tested for a given length of time under some regular or stressed conditions to observe if the units survive or fail the test. The goal is to demonstrate

Dr. Lu is an Assistant Professor in the Department of Mathematics and Statistics. Her email is lulu1@usf.edu.

Dr. Li is an Assistant Professor in the Department of Industrial and Management Systems Engineering. His email is mingyangli@usf.edu.

Dr. Anderson-Cook is a Research Scientist in the Statistical Sciences Group. She is a Fellow of ASQ and the American Statistical Association. Her email is candcook@lanl.gov.

that the reliability that is measured by the probability that a unit survives the test is at or above a certain required level with a desired level of confidence. The determination of a test plan requires specifying a combination of (n, c) values, where n is the number of test units and c is the maximum number of failures allowed for a test to be considered successful. The structure of the test is based on controlling the potential risks associated with making an incorrect decision.

Two types of risks are commonly considered in determining the parameters of a demonstration test plan. One is the *consumer's risk*, which considers the connection between passing the test when the reliability of the unit is not sufficiently good. The other is the *producer's risk*, which focuses on the connection between failing the test when the reliability is good enough. Let π denote the actual reliability at a given time point of interest and let π_0 and π_1 denote the minimum acceptable reliability level and the maximum rejectible reliability level, respectively, where $\pi_1 \leq \pi_0$. The region $\pi \in (\pi_1, \pi_0)$ is called an indifference region (Hamada et al. (2008), p. 344) and can be thought of as the acceptable target region for the true reliability of the unit.

From the frequentist (classical) point of view, the consumer's risk is formally defined as

$$\begin{aligned}\text{FCR} &= P(\text{Test is passed} \mid \pi_1) \\ &= \sum_{y=0}^{c} \binom{n}{y} (1-\pi_1)^y \pi_1^{n-y},\end{aligned}$$

where y denotes the number of observed failures out of n units in the binomial test. Hence, the frequentist consumer's risk (FCR) can be considered as the type-II error probability for passing the test when in fact it should have been failed. Similarly, the frequentist producer's risk (FPR) is defined as

$$\begin{aligned}\text{FPR} &= P(\text{Test is failed} \mid \pi_0) \\ &= \sum_{y=c+1}^{n} \binom{n}{y} (1-\pi_0)^y \pi_0^{n-y}.\end{aligned}$$

It can be considered as the type-I error probability for failing the test when in fact it should have been passed. Easterling (1970) expanded the classical risk criteria by considering the average operating characteristics for a range of acceptable or rejection reliability values. The proposed average risk criteria are defined as follows. The average consumer's risk (ACR) is defined as the probability of passing a test when reliability is actually in the rejection region, i.e., $\pi \leq \pi_1$, which can be obtained by

$$\begin{aligned}\text{ACR} &= P(\text{Test is passed} \mid \pi \leq \pi_1) \\ &= \frac{\displaystyle\int_0^{\pi_1} \left[\sum_{y=0}^{c} \binom{n}{y} (1-\pi)^y \pi^{n-y}\right] p(\pi) d\pi}{\displaystyle\int_0^{\pi_1} p(\pi) d\pi}.\end{aligned}$$

Calculating ACR requires specifying a suitable prior distribution, $p(\pi)$, for reliability π, which represents underlying knowledge about the reliability performance before conducting the test. This is usually identified based on either historical data for similar products or subject matter expert judgement. If no such prior information is available, then a noninformative or diffuse prior distribution can be used. The integration in the above formula can be approximated through a discrete numerical approximation, such as Monte Carlo simulation. The average producer's risk (APR) is defined as the probability of failing the test when reliability is actually in the acceptable region, i.e., $\pi \geq \pi_0$, which can be calculated by

$$\begin{aligned}\text{APR} &= P(\text{Test is failed} \mid \pi \geq \pi_0) \\ &= \frac{\displaystyle\int_{\pi_0}^{1} \left[\sum_{y=c+1}^{n} \binom{n}{y} (1-\pi)^y \pi^{n-y}\right] p(\pi) d\pi}{\displaystyle\int_{\pi_0}^{1} p(\pi) d\pi}.\end{aligned}$$

In recent decades, Bayesian methods have been used more often in reliability analysis and allow the ability to answer a broader range of questions of interest. In the Bayesian framework, the two types of risks can be measured by their corresponding posterior probabilities (Hamada et al. (2008), pp. 346–347). More specifically, the posterior consumer's risk (PCR) is defined as the probability that reliability is in fact in the rejection region given that the test is passed, which is the posterior probability of $\pi \leq \pi_1$ given that no more than c failures have been observed. For a binomial test, the PCR can be obtained by

$$\begin{aligned}\text{PCR} &= P(\pi \leq \pi_1 \mid \text{Test is passed}) \\ &= \frac{\displaystyle\int_0^{\pi_1} \left[\sum_{y=0}^{c} \binom{n}{y} (1-\pi)^y \pi^{n-y}\right] p(\pi) d\pi}{\displaystyle\int_0^{1} \left[\sum_{y=0}^{c} \binom{n}{y} (1-\pi)^y \pi^{n-y}\right] p(\pi) d\pi}. \quad (1)\end{aligned}$$

It should be noted that the conditional probability here has been reversed. In the frequentist version of

the average consumer's risk, the assumption is that the part is not good enough ($\pi \leq \pi_1$) and the risk measures how likely it is that the test is passed. The Bayesian version of the consumer's risk looks at a different summary, which asks about the probability that the part is not good enough, conditioning on seeing a pass on the test. One could argue that this is more likely to be what the consumer is interested in because it presupposes that the test has been passed and the product is being distributed as sufficiently reliable. The posterior producer's risk (PPR) is defined as the posterior probability of $\pi \geq \pi_0$ given that more than c failures have been observed (test is failed), which can be calculated by

$$\begin{aligned}\text{PPR} &= P(\pi \geq \pi_0 \mid \text{Test is failed}) \\ &= \frac{\int_{\pi_0}^{1} \left[\sum_{y=c+1}^{n} \binom{n}{y} (1-\pi)^y \pi^{n-y} \right] p(\pi) d\pi}{\int_{0}^{1} \left[\sum_{y=c+1}^{n} \binom{n}{y} (1-\pi)^y \pi^{n-y} \right] p(\pi) d\pi}. \end{aligned} \quad (2)$$

Again, the conditional probability has been reversed compared with the frequentist version to focus on what is at risk if the test has been deemed a failure.

The different versions of risk criteria allow the practitioners to have more flexibility to choose the most suitable criteria for their applications. For example, if a practitioner has specific desirable and undesirable reliability values in mind, then the classic risk criteria can be better indications of the quantity of real interest. But if the practitioner considers a range of values to be acceptable or unacceptable, then the average risk criteria or the posterior risk criteria can be more suitable metrics for quantifying the risks associated with the range of possible values. The choice between the two depends on which point of view among the frequentist and Bayesian analysis the practitioner is more comfortable taking, as the former measures the conditional probability of observing a certain test result given a certain range of reliability values, while the latter measures the reversed conditional probability of having a certain range of reliability given a certain observed test result (pass or fail). The practitioner should choose the right metric for quantifying their relevant risk probabilities. An additional advantage of using either the average or posterior risk criteria is that they make use of supplementary data and information, such as earlier test results and/or subject expert knowledge. The additional information leveraged from supplementary data or information can be used to reduce the amount of testing required. In our paper, we use the Bayesian risk criteria for evaluating and selecting test plans because it seems to better capture the risks that many consumers and producers would be interested in. More details regarding the computation of the Bayesian risk criteria for a binomial demonstration test are given in Section 2.

Once the risk criteria are chosen, then the demonstration test plans are based on the level of risks the practitioners are willing to accept for their particular applications and given resources. For example, the most popular test plans in current practices are the *zero-failure test* or *success run test* (O'Connor and Kleyner (2012)), in which case a test is passed only if there are no failures observed for all tested units. The number of test units is determined by the minimum n needed to ensure the consumer's risk within an acceptable level. The zero-failure test plans are popular because they minimize cost by testing the smallest number of possible units while controlling the consumer's risk. However, this test plan completely ignores the producer's risk, which has a strong trade-off with the consumer's risk. In other words, simply focusing on reducing consumer's risk leads to the deterioration of the producer's risk and vice versa. Hence, using the zero-failure test plan could force the producer to take on unacceptably large risk by requiring the investment of large amounts of resources and effort to produce unnecessarily highly reliable product to be able to confidently pass the test.

In addition, having too rigorous a test plan is often associated with a low probability of having a successful test. In product development, this leads to extra cost and effort in redesign and retesting of the product. Hence, focusing too much on the cost of the test may lead to an unnecessarily low probability of passing the test. In other applications, it is beneficial to not have too large a demonstration test as the cost of implementing the test might be prohibitive. Therefore, rather than simply adopting the zero-failure test plan, it makes more sense to quantitatively evaluate the actual criteria for different tests and then examine the trade-offs between consumer's and producer's risks, the cost of the test, and how hard it is to pass. With this information available, it is possible to make a balanced decision more tailored to the specific goals of the test and the practitioner's needs.

Given the multiple facets to consider for a demonstration test plan, this paper explores the selection of an optimal plan based on simultaneously balancing these four criteria. The relationship between them

and their impacts on the decision is studied using a case study from Hart (1990) and re-evaluated by Hamada et al (2008, p. 347). In this case study, a binomial test plan (n, c) is sought for a new modem "B", which is similar to an earlier modem "A" that is highly reliable and currently in production. The two modems are built by the same production line and use most of the same components and the main difference is that modem B operates at a different frequency than modem A. Hart (1990) reports a binomial test for modem A on 150 units with six failures. This results in a 0.1 quantile of A's posterior reliability as 0.938, which is adopted as the lowest acceptable reliability for the new modem B. Due to the similarity between the designs of the two modems, Hamada et al. (2008, p. 347) suggested making use of test A data by incorporating this information into the prior distribution. Because the modems are thought to be very similar but not identical, they treat an A test as "worth" 60% of a B test, which is equivalent to treating 150 modem A test units as $150 \times 0.6 = 90$ modem B test units with $6 \times 0.6 = 3.6$ failures and $144 \times 0.6 = 86.4$ successes. Because the beta distribution is the conjugate prior for the binomial distribution (Gelman et al. (2003)), a beta$(a+1, b+1)$ distribution is commonly used in Bayesian reliability analysis for capturing historical data on $a + b$ test units with a observed successes and b failures (Pintar et al. (2012)). Hence, we initially assume a beta$(87.4, 4.6)$ prior distribution for reliability π to summarize the information from the earlier test. However, we do realize the 60% equivalence of test A data is a subjective choice made based on subject expert's opinion about the similarity of units and their potential failure modes, which could vary between different people. Therefore, different prior distributions are explored later to understand their potential impact on the results and decision making.

The remainder of this paper is organized as follows. Section 2 provides more details on the calculation of the criteria values for the consumer's and producer's risks and the probability of accepting the test using a Bayesian approach and includes a brief introduction on the Pareto front approach used for multiple objective optimization. Section 3 explores the relationship between the multiple criteria and their trade-offs using the case study example from Hart (1990). Three decision-making strategies based on considering multiple criteria simultaneously for different user priorities are outlined and illustrated. Then sensitivity analyses to evaluate the impact of different subjective user choices, such as the threshold value for capping the consumer's risk and the use of different prior distributions, are discussed in Section 4. Section 5 contains some concluding remarks.

2. Optimization Criteria and the Pareto Front Approach

In this section, we present background on several building blocks required for selecting a best demonstration test plan while simultaneously balancing multiple objectives using the Pareto front approach. The first part gives detailed information on calculating the different criteria values including the consumer's and producer's risks and the probability of accepting the test. The second part provides some background on multiple objective optimization using a Pareto front approach.

2.1. Bayesian Posterior Risks

In this paper, we choose to use the Bayesian posterior risks defined in Equations (1) and (2) for quantifying the risk criteria of interest for a demonstration test because we feel the posterior risks are of more direct interest to many consumers and producers because they measure their specific risks based on having observed test results. In addition, they allow us to incorporate earlier test data results. Based on the adjusted number of observed successes (86.4) and failures (3.6), we construct a prior distribution of $\pi \sim \text{Beta}(87.4, 4.6)$ from the modem A test data. Using this, we obtain a large number, say $M = 1{,}000$, draws of possible π values from its prior distribution. Suppose the jth draw is denoted by $\pi^{(j)}$, where $j = 1, \ldots, M$. Then we can evaluate the posterior risks by using Monte Carlo integration based on the M simulated samples from the prior distribution. More specifically, we can approximate the posterior consumer's risk by

$$\begin{aligned}\text{PCR} &= P(\pi \le \pi_1 \mid \text{Test is passed}) \\ &= \left\{ \sum_{j=1}^{M} \left[\sum_{y=0}^{c} \binom{n}{y} (1-\pi^{(j)})^y (\pi^{(j)})^{n-y} \right] \right. \\ &\qquad \left. \times I(\pi^{(j)} \le \pi_1) \right\} \\ &\quad \div \left\{ \sum_{j=1}^{M} \left[\sum_{y=0}^{c} \binom{n}{y} (1-\pi^{(j)})^y (\pi^{(j)})^{n-y} \right] \right\}. \end{aligned} \tag{3}$$

Similarly, the posterior producer's risk can be approximated by

$$\begin{aligned}\text{PPR} &= P(\pi \geq \pi_1 \mid \text{Test is failed}) \\ &= \left\{ \sum_{j=1}^{M} \left[1 - \sum_{y=0}^{c} \binom{n}{y} (1-\pi^{(j)})^y (\pi^{(j)})^{n-y} \right] \right. \\ &\qquad \left. \times I(\pi^{(j)} \geq \pi_1) \right\} \\ &\quad \div \left\{ \sum_{j=1}^{M} \left[1 - \sum_{y=0}^{c} \binom{n}{y} (1-\pi^{(j)})^y (\pi^{(j)})^{n-y} \right] \right\}. \end{aligned} \tag{4}$$

The probability of accepting the design, called the acceptance probability (AP), is the probability of passing the test and can be estimated approximately using

$$\begin{aligned}\text{AP} &= P(\text{Test is passed}) \\ &= \frac{1}{M} \sum_{j=1}^{M} \left[\sum_{y=0}^{c} \binom{n}{y} (1-\pi^{(j)})^y (\pi^{(j)})^{n-y} \right]. \end{aligned} \tag{5}$$

Note that, even though we choose to use Bayesian criteria for our case study, the general methodologies for considering and balancing multiple objectives can be easily adapted for a variety of metrics, including the frequentist classical or average risks, selected as the quantitative risk criteria. For simplicity of discussion below, we abbreviate the notation and use CR and PR to represent the consumer's risk and the producer's risk in the remainer of the paper (including tables and plots), even though, for this case study, they are the posterior risks calculated with Equations (3) and (4). Also, $M = 1000$ was chosen in our example for fast computing for demonstrating the proposed methods. However, slightly more precise approximations of the criteria values could be achieved with a larger M value.

2.2. Pareto Front Optimization

We now provide a brief background on multiple objective optimization using a Pareto front approach. Optimal decision making based on multiple objectives or responses has been receiving more attention recently in many fields due to the increasingly constrained budgets and resources. In many applications, different objectives compete for resources when considered simultaneously. The "ideal" solution of attaining the best possible outcome for all criteria usually does not exist. Hence, because some compromise is needed, it is important to understand the trade-offs between the objectives to make a balanced decision to match the goals of a particular study.

Constrained optimization (Lange (2013), p. 10) and the desirability function approach (Derringer and Suich (1980)) were among the most popular tools used for finding a single "best" solution for optimizing multiple objectives until the Pareto front approach (Kasprzak and Lewis (2001), Gronwald et al. (2008), Trautmann and Mehnen (2009), Lu et al. (2011)) came to be used more extensively to find a set of competing solutions for all objectives under consideration. A feasible solution is said to *Pareto dominate* another solution if it is as good as the other solution based on all objectives and is strictly better for at least one of the objectives. The *Pareto optimal set* contains all nondominated solutions that cannot improve any of the objectives without deteriorating one of the other objectives and hence provides a complete set of superior solutions from which a rational final choice should be selected. The corresponding objective function values of all solutions in a Pareto optimal set form a *Pareto front* in the criteria space. For problems with an infinite or extremely large solution space that is infeasible to enumerate, some search algorithms are needed to efficiently populate the Pareto front. Compared with mathematical programming-based methods, the evolutionary algorithms (Deb (2009)) based on applying Pareto ranking to solutions has become more popular recently to generate Pareto optimal solutions for multiple response optimization problems. Lu et al. (2011) further developed the Pareto front approach into a structured two-stage decision-making process: the first stage is an objective stage that identifies superior choices (the Pareto optimal set) and eliminates noncontenders from future consideration, while the second stage examines the identified choices and selects which one(s) is (are) best to match the specific goals of the study. The overall approach matches the define-measure-reduce-combine-select process described in Anderson-Cook and Lu (2015).

In this paper, the Pareto front approach is used to find a collection of nondominating test plans based on considering multiple criteria simultaneously to help eliminate noncontenders from decision making. Due to the relative small size of the problem for specifying practically realistic demonstration tests, the Pareto optimal solutions are identified from an enumerated set of possible tests for specified ranges of (n, c) values based on directly applying Pareto ranking to all solutions under consideration. If a larger scale test becomes possible for certain applications, then a tailored evolutionary search algorithm could

be used to more efficiently populate the Pareto front. For the case study, once the superior choices have been identified, we then highlight some graphical summaries that can help the practitioner determine which demonstration test is best suited to the particular needs of the study.

3. Case Study

This section illustrates the decision-making process for choosing a best demonstration test plan based on simultaneously considering multiple objectives using the case study from Hart (1990) introduced in Section 1. Four criteria described in Sections 1 and 2, including the consumer's risk (CR = PCR), producer's risk (PR = PPR), acceptance probability (AP), and cost, which is measured by the number of test units (n), are considered. To fully examine the relationship between the four criteria with the test plan parameters (n, c), an exhaustive evaluation of all possible test plans for a range of possible parameters with $c \in [0, 20]$ and $n \in [c + 1, 500]$ was conducted. Larger (n, c) values are not examined because these tests are not considered practical for many real applications. However, the methodology and the general conclusions can be easily adapted to evaluate any set of tests under consideration.

There are 10,290 test plans $[(n, c) = (1, 0), (2, 0), \ldots, (500, 0), (2, 1), (3, 1), \ldots, (500, 1), (3, 2), \ldots]$ evaluated in this case study. For each test plan (n, c), the four criteria values are calculated using the formulas given in Equations (3)–(5) plus the number of test units n. Before starting the selection process for a particular scenario, it is helpful to investigate the relationship between these criteria to gain some intuition about the degree of trade-off between them and how strongly correlated they are. Figure 1 shows several snapshots highlighting different pairs of criteria to illustrate interrelationships between the four criteria. Figure 1(a) is a plot of CR vs. PR for all the examined test plans, which displays a set of curves in different gray shades for different c values with darker colors representing smaller c values. A few prominent patterns can be observed from this plot. First, for each fixed c (corresponding to a single curve), there is a strong trade-off between CR and PR because, as one risk improves, the other gets worse. Second, as we increase the maximum allowable failures (c), both CR and PR can be simultaneously improved by increasing n, which is evidenced by the progression of gray shades becoming lighter as the curves moving closer to the ideal bottom left corner (corresponding to no risk for either producer or consumer). However, there are diminishing returns as c increases. Third, the range of CR is bounded between 0 and 0.25, while the range of PR is substantially wider between 0 and around 0.75. When the popular zero-failure test plan is used ($c = 0$), the minimal PR is around 0.59, which is considerably higher than any typically acceptable standard, and this is achieved when testing a single unit with CR around 0.25. As we increase n to reduce CR, the PR becomes even larger. When CR is controlled to be around 0.1, the PR is above 0.68. One of the big drawbacks of using the minimum sample size test plan is that it offers no protection from the producer's risk, which means that a producer is faced with a difficult option—produce units with reliability well above the required range of reliability or bear considerable risk of failing the test. In addition to the curves, symbols shows some specific test plans with $n = 50$, 100, 200, 300, and 400 to highlight the effects of changing the sample size. We can see that, first, for any fixed n, we can reduce the CR by allowing fewer maximum failures (c). However, this results in a quick rise in PR. On the other hand, allowing more maximum failures reduces the PR but increases the CR. Second, for any fixed maximum allowable failures c, increasing n reduces the CR but increases the PR, and vice versa. There is also diminishing effect on the amount of risk reduction possible by adjusting the sample size while sacrificing the other risk.

Figure 1(b) plots AP vs. sample size n for all of the test plans with different c values (dark to light gray, indicating small to large c values). The following patterns can be observed: First, for each fixed c value, the probability of passing the test drops as sample size n increases. This is because, for fixed maximum allowable failures, the more units we test, the smaller chance there is to pass the test. Second, for a fixed sample size n, the chance of accepting the test increases as more maximum failures are allowed. In addition to the curves with different c values, test plans achieving controlled CR levels at 0, 0.02, 0.04, 0.06, 0.08, and 0.1 and test plans with controlled PR levels at 0.1, 0.2, 0.3, 0.4, and 0.5 are highlighted with different symbols. We can see that, for fixed n, higher AP is generally associated with larger CR, but smaller PR. This is intuitive because, as we allow fewer maximum failures to reduce CR for a fixed sample size, passing the test becomes less probable and the risk for the producer increases. Meanwhile, if we control the CR at a fixed level, we can increase the probability of passing the test by increasing the sample size n and/or allowing more maximum failures c. Alternately, if we control the level of PR, reducing

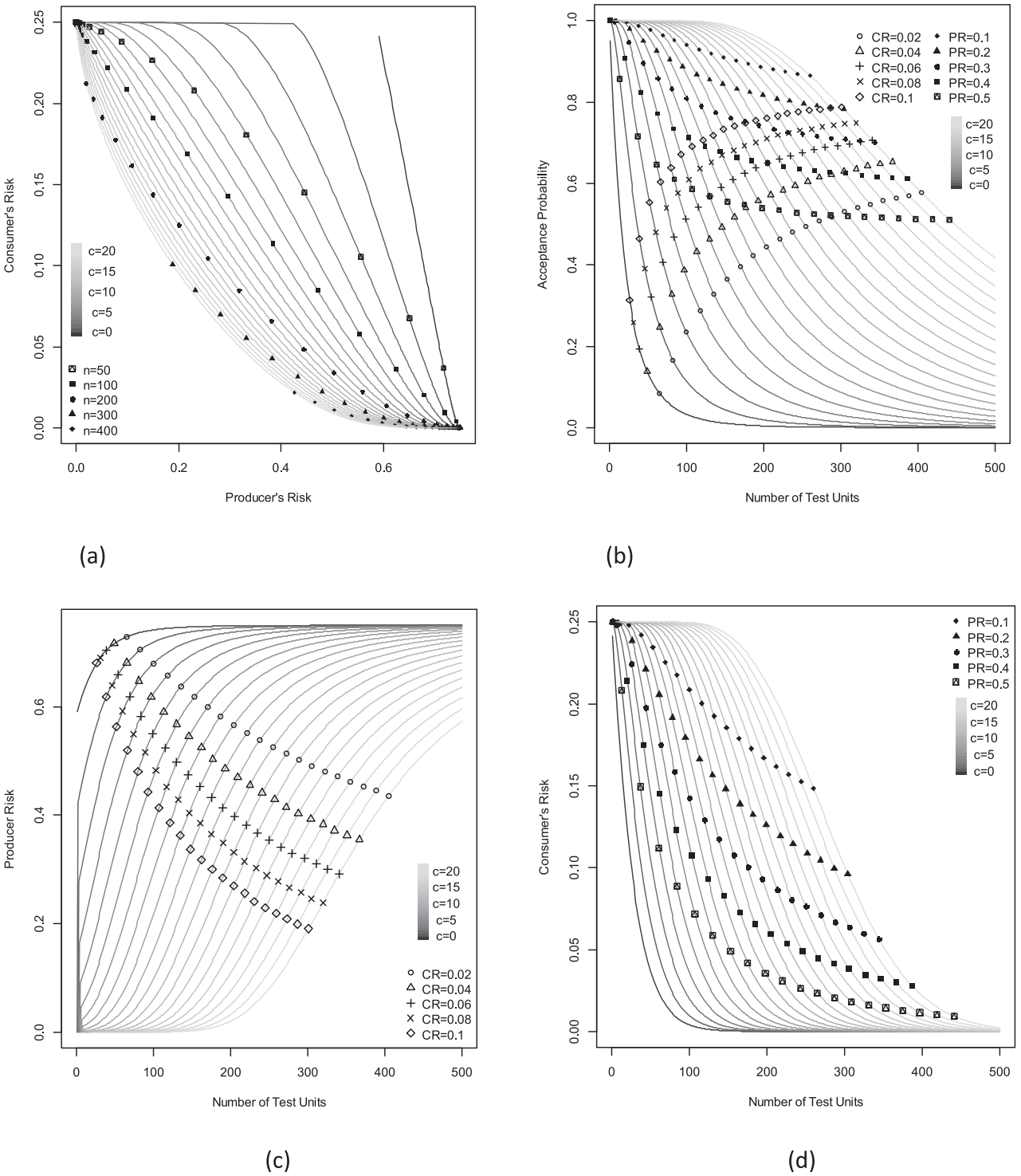

FIGURE 1. Plots Showing the Interrelationships Between the Consumer's Risk (CR), the Producer's Risk (PR), the Acceptance Probability (AP), and the Sample Size n. In each panel, test plans with the same c value are on the same curve and darker to lighter gray shades are used for smaller to larger c values in [0, 20]. Different symbols indicate some representative levels for other criteria: for example, different symbols are used to display different n values in (a) and different controlled risk levels in (b)–(d).

the sample size and maximum allowable failures increases the AP.

Figure 1(c) shows the plot of PR vs. sample size n for different c values, while highlighting some test plans with controlled CR levels at 0, 0.02, 0.04, 0.06, 0.08, and 0.1. Similar information as discussed for other subfigures can be obtained from the plot. For example, the PR increases as more units are tested for fixed c, while the CR is simultaneously reduced. For a fixed sample size, we can reduce the PR by allowing more failures, while increasing CR. If we control the CR at a fixed level, we can reduce the PR by simultaneously increasing n and c. Similarly, Figure 1(d) shows the plot of CR vs. sample size n for different c values, while highlighting some test plans with controlled PR levels at 0.1, 0.2, 0.3, 0.4, and 0.5. Again, the CR can be reduced and PR is increased as more units are tested with fixed maximum allowable failures. For fixed sample size, the CR is reduced and PR increases by allowing fewer failures. If the PR is controlled at a fixed level, then we can also reduce the CR by simultaneously increasing n and c.

As a summary, the CR and PR have the most trade-off among all criteria under consideration. When one of n or c are fixed, we can adjust the other parameter to reduce one of the risks, but it also raises the other risk. The only way to reduce both types of risks is to simultaneously increase n and c. However, this increases the cost of the test and possibly reduces the chance of passing the test. Hence, as is often the case when considering multiple objectives, there is no universal solution to simultaneously optimize all criteria under consideration. In order to select the best test for a given scenario, the practitioner needs to prioritize the competing objectives and make a tailored decision to best match their study goals. Below we present three possible paths to guide test selection based on different user priorities. The strategies begin by constructing the Pareto front of nondominating solutions after initially controlling one of the criterion values. From there, it is possible to examine the remaining choices in a more manageable form and select one that is well suited to the demonstration test goals.

The first strategy considers when the practitioner determines that the consumer's risk is most important among all criteria. This is not unusual considering the majority of current demonstration tests focus solely on controlling CR without actively considering other aspects of their decision. Assume that we wish to bound the consumer's risk at no more than 0.2. Then among all the test plans with acceptable CR, we can construct the Pareto front with the set of nondominated solutions based on the remaining three criteria. Figure 2 shows the PR, AP, and n for all the test plans on the Pareto front given the upper bound of CR chosen at 0.2. The most prominent feature of Figure 2 is the tremendous simplification that is achieved with this constraint. This Pareto front contains only 21 test plans with a single test plan for each different c value. This indicates, for each fixed c value, there is a universal optimal test plan when simultaneously considering PR, AP, and n given the constraint on CR. This property allows the practitioner to quickly reduce their options to a manageable number and then choose a single best test plan based on their goals for the other three criteria. Figure 2 shows the trade-offs between the three criteria given the constraint on CR. The test plans are sorted from left to right with increasing c value from 0 to 20. The left vertical axis is scaled between 0 and 1. The PR and AP are measured on a probability scale, which is labeled on the left axis. For cost measure with the number of test units n, the right axis values are scaled between 0 and 250 to include all test plans on the Pareto front. This plot, called the trade-off plot (Lu et al. (2011)), is effective in showing the amount of trade-offs between competing solutions. The sample size increases from 7 for $c = 0$ to 214 for $c = 20$, the PR drops from around 0.6 to below 0.05, and the AP increases from 0.7 to close to 0.95. As we increase n and c, we can simultaneously reduce PR and improve AP. However, the rate of improvement diminishes as n and c increase. The criteria values are listed in the first four columns of Table 1. With this trade-off plot, users can make their own tailored decision based on the available budget and time, the level of risk they can tolerate, or the lowest probability acceptable for a successful test. For example, if the practitioner can tolerate up to 0.2 level of PR, then the best plan is to test 74 units allowing up to 7 failures for a successful test. With this plan, there is a 0.92 probability of accepting the design while controlling both CR and PR to be no larger than 0.2. However, if the practitioner has a tight budget and can afford no more than 50 units, then the best plan is to test 44 units while allowing no more than 4 failures. With this plan, there is nearly a 0.3 producer's risk and nearly a 0.9 probability to pass the test.

The second strategy prioritizes the producer's risk as most important among all criteria. Consider that we want the PR at or below the 0.2 level. The Pareto front of all contending solutions considering the re-

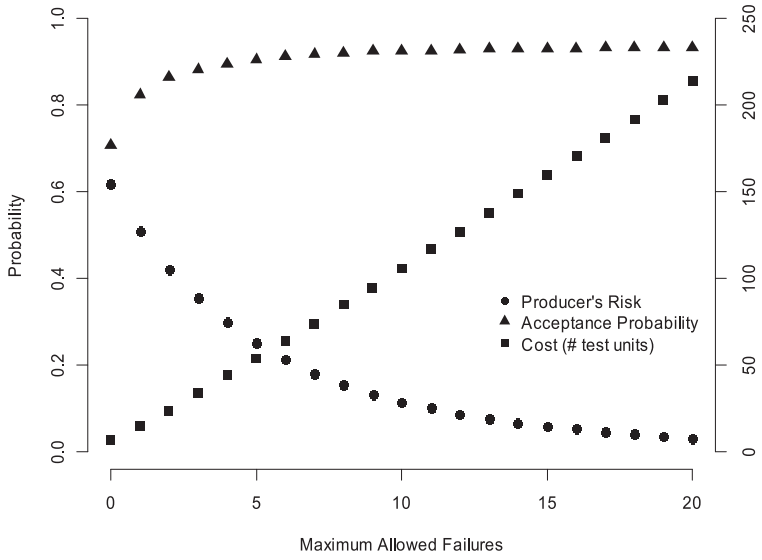

FIGURE 2. Trade-Off Plot for the 21 Choices on the Three Criteria Pareto Front Based on Producer's Risk, Acceptance Probability and Cost, Given the Constraint That Consumer's Risk Is No Greater than 0.2. Note each choice corresponds to a different c value ranging between 0 and 20, sorted from left to right with increasing c values. The left-axis scale gives the producer's risk (PR) and acceptance probability (AP). The scale on the right axis shows the number of test units n.

TABLE 1. Summary of 21 Options on the Pareto Front Based on Producer's Risk (PR), Acceptance Probability (AP), and the Number of Test Units (n) with the Upper Bound for the Consumer's Risk Set at 0.20, 0.10, and 0.05 when A Test Data Are Considered to Be 60% Equivalent to B Test Data. Note that there is a unique optimal test plan for each choice of c

	Consumer's risk (CR) upper bound								
	0.2			0.1			0.05		
c	PR	AP	n	PR	AP	n	PR	AP	n
0	0.6192	0.7087	7	0.6803	0.3131	26	0.7095	0.1679	43
1	0.5099	0.8237	15	0.6183	0.4646	39	0.6701	0.2844	59
2	0.4221	0.8661	24	0.5636	0.5536	52	0.6343	0.3646	75
3	0.3546	0.8839	34	0.5193	0.6037	66	0.5998	0.4284	90
4	0.2984	0.8967	44	0.4809	0.6388	80	0.5712	0.4709	106
5	0.2516	0.9062	54	0.4429	0.6711	93	0.5427	0.5089	121
6	0.2125	0.9135	64	0.4130	0.6905	107	0.5197	0.5345	137
7	0.1798	0.9193	74	0.3865	0.7059	121	0.4989	0.5552	153
8	0.1562	0.9208	85	0.3626	0.7185	135	0.4772	0.5764	168
9	0.1327	0.9250	95	0.3374	0.7330	148	0.4601	0.5905	184
10	0.1160	0.9260	106	0.3180	0.7415	162	0.4415	0.6061	199
11	0.1016	0.9269	117	0.3005	0.7488	176	0.4242	0.6195	214
12	0.0868	0.9298	127	0.2844	0.7551	190	0.4111	0.6282	230
13	0.0764	0.9303	138	0.2697	0.7607	204	0.3961	0.6387	245
14	0.0674	0.9309	149	0.2561	0.7655	218	0.3848	0.6452	261
15	0.0596	0.9313	160	0.2406	0.7725	231	0.3717	0.6537	276
16	0.0528	0.9318	171	0.2291	0.7762	245	0.3593	0.6613	291
17	0.0455	0.9337	181	0.2185	0.7796	259	0.3502	0.6658	307
18	0.0404	0.9339	192	0.2086	0.7826	273	0.3392	0.6722	322
19	0.0359	0.9342	203	0.1994	0.7854	287	0.3288	0.6780	337
20	0.0320	0.9345	214	0.1908	0.7880	301	0.3214	0.6812	353

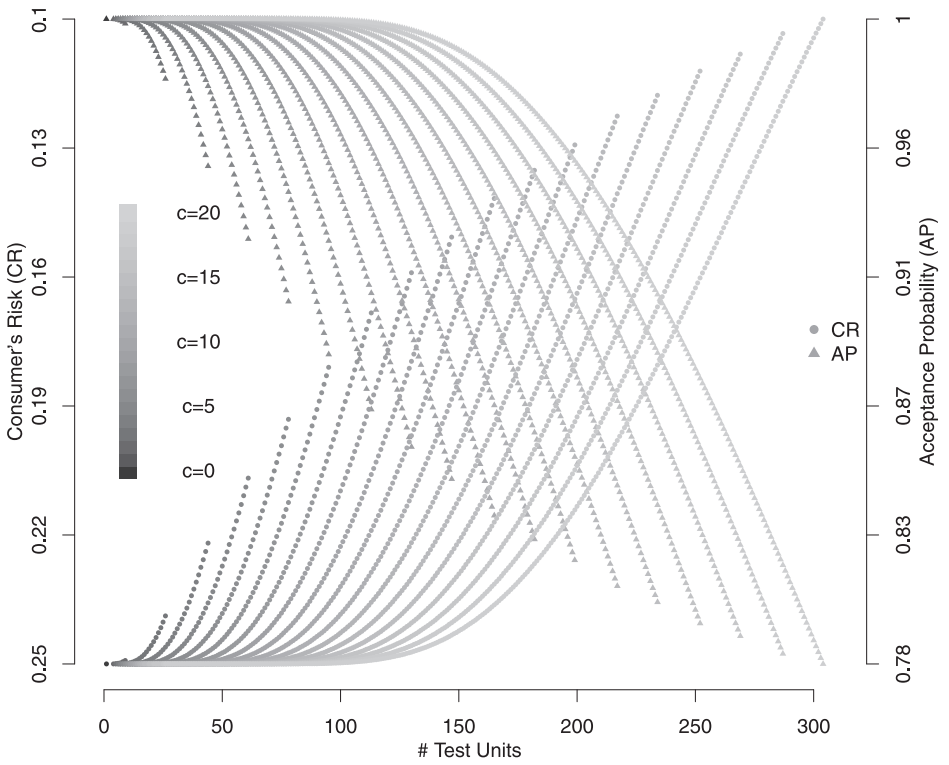

FIGURE 3. Trade-Off Plot for the 2592 Test Plans on the Pareto Front Based on the Consumer's Risk, the Acceptance Probability and the Cost, Given a Producer's Risk Is No Greater than 0.2. Different shades of gray colors distinguish test plans with different c values, with darker gray for smaller c values. The left-axis scale gives consumer's risk and the right-axis scale is for the acceptance probability. Both scales display the best values at the top (the minimum value for the consumer's risk and the maximum value for the acceptance probability) and the worst values at the bottom.

maining three criteria is shown in Figure 3. In this figure, the horizontal axis shows the number of test units (i.e., the cost). For each test plan of a certain sample size, the diamond and triangle symbols are used to display its CR and AP values, respectively. The left axis shows the CR scale with the best value (the minimum risk for all the choices on the front) on the top and the worst value (the maximum risk) at the bottom. The right axis shows the range of the AP for test plans on the Pareto front, with also the best value (the maximum AP) shown on the top and worst value (the minimum AP) at the bottom. Note, for this plot, the ideal values for each criterion are scaled to be at the top of the figure. Similar to Figure 1, the darker to lighter gray shades are used to identify smaller to larger c values.

A key pattern highlighted in Figure 3 is how much richer the set of nondominated choices is when controlling the PR with an upper bound compared with controlling the CR. The Pareto front based on the three criteria other than the PR contains 2592 test plans, with a maximum sample size of 310 units, the CR between 0.1 and 0.25 and the AP above 0.78. There are many choices for each possible c value. This indicates that, given a fixed number of maximum allowable failures, there are trade-offs between cost and the other two criteria. Particularly, as we increase the sample size, we see incremental improvements in the CR with simultaneous reduction in the AP. Also, as more failures are allowed, there are more competing options with more largely varied sample size, which allows even more trade-offs between the competing CR and AP criteria. Due to the richness of the Pareto front, further decision making is less straightforward compared with controlling CR primarily. To proceed, the user should prioritize the remaining criteria to reduce the number of options to a manageable number. For example, based on cost and logistical constraints, there might be a limited number of testing devices available, say less than 100, which can reduce the set of options to choose between based on CR and AP to identify minimum (n, c) values to accommodate the requirements on both criteria.

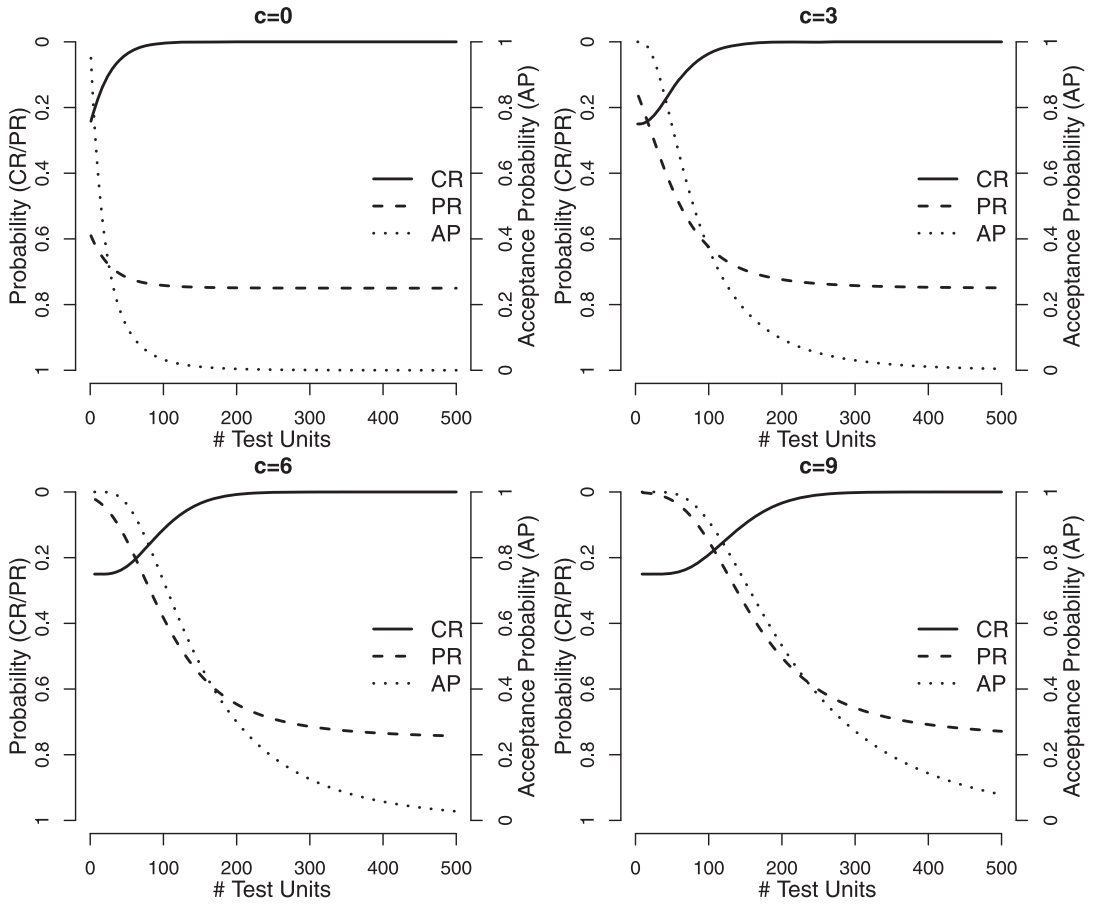

FIGURE 4. Trade-Off Plot for Test Plans on the Four Criteria Pareto Front for Fixed c Values (Maximum Allowable Failures) at $c = 0$, 3, 6, and 9. For each panel, the left axis is for the consumer's and producer's risks while the right axis is for the acceptance probability. Note that, for all criteria, the best values are displayed at the top.

A third strategy involves the user specifying a particular c value for the test. Figure 4 shows the trade-off plot for all test plans on the four criteria Pareto front for specific fixed c values of 0, 3, 6, and 9. For each panel with a fixed c, the horizontal axis shows the sample size. The CR, PR, and AP values for all contending test plans on the front are shown in different line types across the range of sample size for $n \in [1, 500]$. The left axis provides the CR and PR scales with the best value (minimum risk at 0) at the top and the worst value (maximum risk at 1) at the bottom. The right axis is for the AP scale with the best value (maximum acceptance probability at 1) on the top and worst value (0) on the bottom. Therefore, the closer a curve is to the top of the plot, the better the performance for the corresponding criterion.

Consider the top left panel for example. This plot corresponds to the commonly used zero-failure test. As more units are tested, the CR quickly improves. But the trade-off is that both AP and PR become worse. The AP drops quickly to below 0.2 as sample size increase to around 50 and CR is reduced to around 0.1. The PR increases from around 0.6 to 0.7, both of which are extremely high from the producer's perspective. However, the changes diminish when n is increased beyond 100 units. If $c = 3$ is selected, then a plan with around 80 test units allows around 0.1 for CR, 0.5 for AP, and slightly above 0.5 for PR. Similar patterns are observed from the bottom two panels, with larger c values encouraging improvement in PR and AP by increasing the sample size to achieve a more acceptable balance between the CR and those two criteria.

As a summary, we have illustrated three strategies of decision-making when choosing a best demonstration test plan. Which strategy to choose and which final plan is selected depends on the users' priorities as well as the available resources and logistic con-

straints for their particular applications. However, the complexity associated with the three decision-making processes are quite different. Primarily controlling the CR leads to the simplest choices and is most straightforward for reaching a final decision. Focusing on a particular c value leads to a rich set of contending options due to the large amount of trade-offs between all four criteria under consideration. However, there is a simple and clear pattern of interrelationships between the criteria, and the trade-off plot shown in Figure 4 is effective in capturing this pattern and helping make an informed decision. Compared with the above two scenarios, primarily controlling the PR could lead to an overwhelmingly rich set of options. However, using the Pareto front approach to eliminate inferior solutions (from 10,290 to 2592) and using tailored graphical summaries such as the trade-off plot in Figure 3 can be effective for extracting key patterns and helping guide the remaining decision making in a more structured and justifiable way.

4. Sensitivity Analysis

This section explores the impacts of some of the subjective user choices on the decision. We focus our exploration by following the first decision-making path outlined in Section 3 due to its broad applicability and simplicity of implementation. However, this type of study could be conducted for any of the three strategies.

The first user input that could have substantial impact on the set of superior options to consider is the threshold value (upper bound) we use to control the consumer's risk level. Depending on if the user employs a stringent or liberal constraint on the consumer's risk, we could end up with very different sets of superior options from which to choose. Figure 5 shows the trade-off plot for using three different cut-off values for the CR at 0.05, 0.1, and 0.2 levels. For each constraint level, there are three curves representing the criteria values for all test plans on the Pareto front based on the PR, AP, and cost criteria. Note there is again a single best choice for

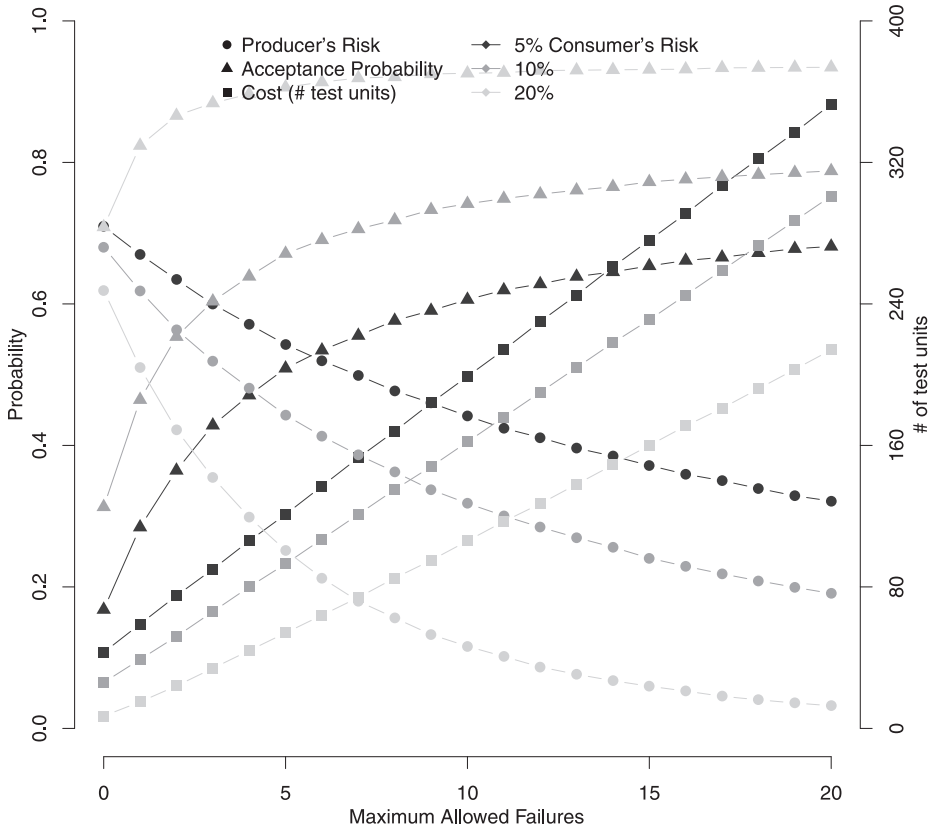

FIGURE 5. Trade-Off Plot for Test Plans on the Four Criteria Pareto Front for Fixed c Values (Maximum Allowable Failures) at $c = 0$, 3, 6, and 9. For each panel, the left axis is for the consumer's and producer's risks while the right axis is for the acceptance probability. Note that, for all criteria, the best values are displayed at the top.

each c value. Three dark-to-light gray shades are used to represent the three stringent-to-liberal constraint levels on the CR. Table 1 shows the actual criteria values for all 21 superior choices with different c values for the three selected threshold levels for CR.

Some observations can be made from Figure 5 and Table 1. First, using a smaller threshold value for the CR requires testing more units for a fixed c value. This also results in considerably higher PR and lower AP values, and the amount of changes in PR and AP values initially increase quickly as c increases but start to diminish as c reaches a certain value. For example, if the user can tolerate no more than 0.5 for producer's risk, then with the use of 0.2 threshold for CR, only 24 units need to be tested to achieve 0.42 for PR and 0.87 for AP. However, with the more stringent constraint on CR, the user needs to increase the number of units tested to 80 to get 0.48 for PR and only 0.64 for AP for using 0.1 threshold, or test 153 units to get 0.50 for PR and only 0.56 for AP for a 0.05 CR cut-off value. Second, considering the diminishing effect of increasing sample size on simultaneously improving PR and AP, if too harsh a standard is used for the CR, then it may be impossible to meet reasonable standards on either PR or AP regardless of the number of units tested. For example, if a 0.05 threshold is used for CR, then, from Figure 5, achieving an AP above 0.80 or a PR below 0.2 is not possible for $n \in [c+1, 500]$. Therefore, it is necessary to evaluate the test performance with different levels of constraint on CR to understand the severity of its impact on the available choices.

When using a Bayesian approach for quantifying the different criteria, the calculated risk criteria and probability of accepting the test can be sensitive to the user-specified prior distribution. Recall that results in Section 3 were based on treating the A test data as 60% equivalent to B test data, which was determined by the subject matter expert based on the similarity of the designs for the two modems. In reality, this is just an approximation and different subject matter experts are likely to have different opinions regarding the relevance and value of the historical data. For example, another subject matter expert may consider A test data as only 40% equivalent to B test data. In this case, the 150 modem A test units are treated as $150 \times 0.4 = 60$ modem B test units with $6 \times 0.4 = 2.4$ failures and $144 \times 0.4 = 57.6$ successes. Hence, a prior distribution of the form of $\pi \sim \text{Beta}(58.6, 3.4)$ should be used. To study the impact of this subjective choice on the amount of information leveraged from the earlier test data, we consider reducing the amount of equivalent information from A test data down to 40% and 20%. Figure 6 shows the probability density curves for the three prior distributions specified based on using relationships of 60%, 40%, and 20% between the current study and the A test data. The center of the three prior distributions are quite similar, while the spread of the distribution increases as less information from earlier test data is leveraged. The lower end of the possible reliability values drops from around 0.85 to 0.8 and further down to 0.75 when the amount of information borrowed from A test data is reduced from 60% to 40% and then to 20%. This matches our intuition that using more historical data can provide stronger evidence of an assumed high reliability for the new modems in our case study.

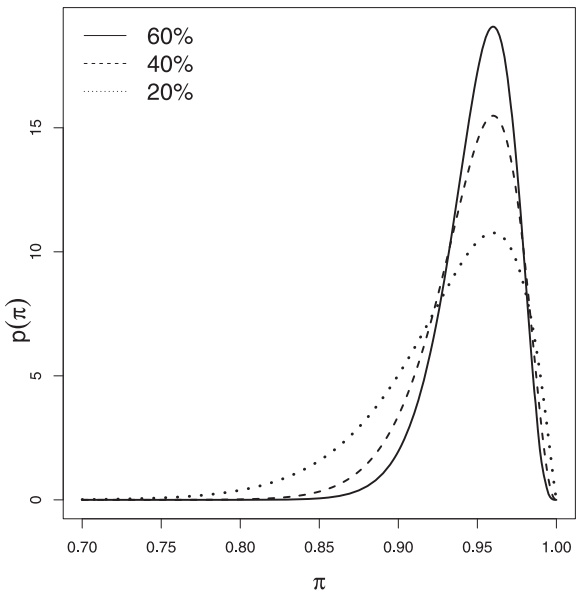

FIGURE 6. Probability Density Curves for Prior Distributions Based on Leveraging Different Amounts (60%, 40%, and 20%) of the A Test Data, Which Result in the Use of Beta(87.4, 4.6), Beta(58.6, 3.4), and Beta(29.8,2.2) Prior Distributions, Respectively.

Figure 7 explores sensitivity to the amount of prior information incorporated from historical data. Figure 7(a) shows the trade-off plot with different constraint levels on CR when using weaker prior information (A test data only 40% equivalent to B test data). Compared with Figure 5, testing more units is required to achieve the same levels on CR with a fixed c. Meanwhile, the corresponding PR is slightly reduced, while AP drops considerably. Figure 7(b)

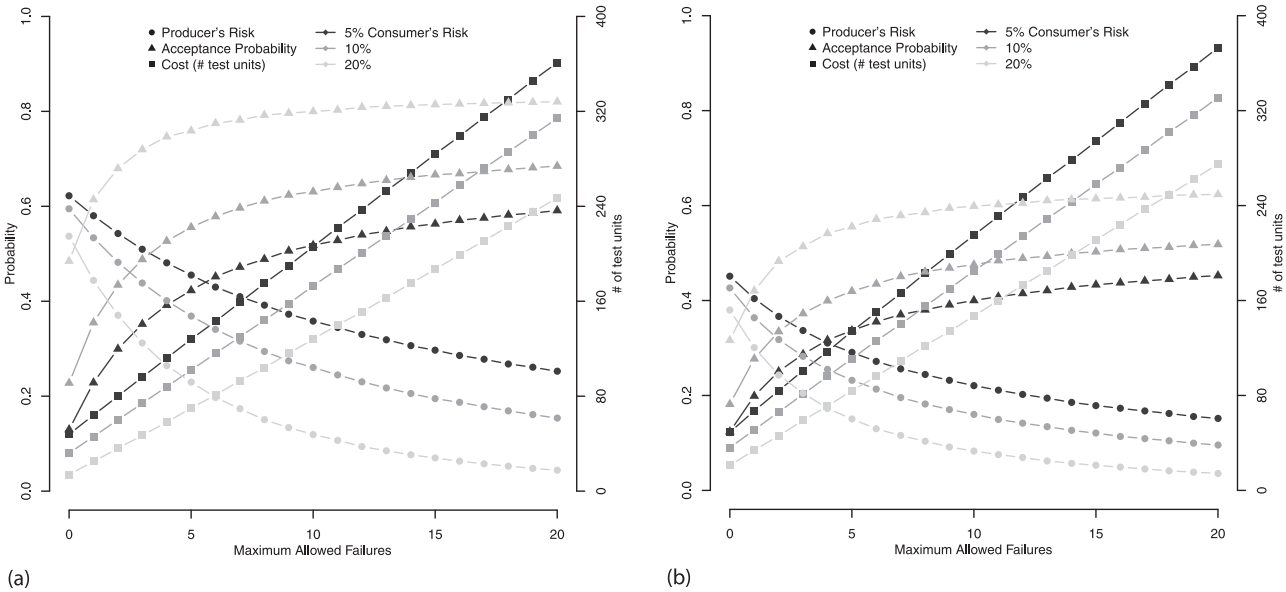

FIGURE 7. Trade-Off Plots Based on Different Prior Distributions. (a) Considering A test data as 40% equivalent to B test data; (b) Considering A test data as 20% equivalent to B test data.

TABLE 2. Summary of 21 Options on the Pareto Front Based on Producer's Risk (PR), Acceptance Probability (AP), and the Number of Test Units (n) with the Upper Bound for the Consumer's Risk Set at 0.20, 0.10, and 0.05 when A Test Data Are Considered to Be 40% Equivalent to B Test Data

	Consumer's risk (CR) upper bound								
	0.2			0.1			0.05		
c	PR	AP	n	PR	AP	n	PR	AP	n
1	0.4438	0.6143	25	0.5337	0.3550	46	0.5802	0.2285	64
2	0.3706	0.6797	36	0.4822	0.4343	60	0.5427	0.2996	80
3	0.3119	0.7197	47	0.4386	0.4881	74	0.5097	0.3519	96
4	0.2643	0.7466	58	0.4012	0.5268	88	0.4808	0.3917	112
5	0.2299	0.7593	70	0.3689	0.5559	102	0.4553	0.4228	128
6	0.1971	0.7748	81	0.3407	0.5785	116	0.4299	0.4519	143
7	0.1737	0.7818	93	0.3160	0.5967	130	0.4098	0.4720	159
8	0.1503	0.7920	104	0.2942	0.6116	144	0.3918	0.4887	175
9	0.1337	0.7963	116	0.2748	0.6239	158	0.3728	0.5062	190
10	0.1194	0.7999	128	0.2605	0.6308	173	0.3582	0.5180	206
11	0.1071	0.8029	140	0.2447	0.6400	187	0.3448	0.5283	222
12	0.0940	0.8088	151	0.2304	0.6480	201	0.3301	0.5400	237
13	0.0848	0.8108	163	0.2174	0.6550	215	0.3189	0.5477	253
14	0.0767	0.8127	175	0.2055	0.6611	229	0.3063	0.5571	268
15	0.0696	0.8143	187	0.1947	0.6666	243	0.2968	0.5631	284
16	0.0632	0.8158	199	0.1871	0.6690	258	0.2858	0.5707	299
17	0.0576	0.8171	211	0.1778	0.6736	272	0.2777	0.5755	315
18	0.0526	0.8182	223	0.1692	0.6777	286	0.2680	0.5819	330
19	0.0481	0.8193	235	0.1612	0.6814	300	0.2610	0.5857	346
20	0.0440	0.8203	247	0.1538	0.6848	314	0.2524	0.5911	361

TABLE 3. Summary of 21 Options on the Pareto Front Based on Producer's Risk (PR), Acceptance Probability (AP), and the Number of Test Units (n) with the Upper Bound for the Consumer's Risk Set at 0.20, 0.10, and 0.05 when A Test Data Are Considered to Be 20% Equivalent to B Test Data

	Consumer's risk (CR) upper bound								
	0.2			0.1			0.05		
c	PR	AP	n	PR	AP	n	PR	AP	n
0	0.3800	0.3163	21	0.4264	0.1816	36	0.4513	0.1238	49
1	0.3008	0.4205	34	0.3638	0.2773	51	0.4042	0.1988	67
2	0.2431	0.4829	46	0.3177	0.3347	66	0.3665	0.2506	84
3	0.2049	0.5138	59	0.2825	0.3726	81	0.3366	0.2869	101
4	0.1726	0.5414	71	0.2547	0.3995	96	0.3103	0.3167	117
5	0.1505	0.5555	84	0.2322	0.4195	111	0.2906	0.3367	134
6	0.1299	0.5714	96	0.2135	0.4350	126	0.2719	0.3553	150
7	0.1156	0.5793	109	0.1953	0.4508	140	0.2559	0.3703	166
8	0.1036	0.5855	122	0.1820	0.4606	155	0.2440	0.3803	183
9	0.0914	0.5947	134	0.1705	0.4687	170	0.2317	0.3910	199
10	0.0829	0.5987	147	0.1604	0.4756	185	0.2208	0.4002	215
11	0.0756	0.6022	160	0.1494	0.4841	199	0.2110	0.4081	231
12	0.0691	0.6052	173	0.1415	0.4890	214	0.2023	0.4150	247
13	0.0621	0.6107	185	0.1345	0.4933	229	0.1943	0.4211	263
14	0.0572	0.6128	198	0.1264	0.4994	243	0.1854	0.4284	278
15	0.0528	0.6147	211	0.1206	0.5027	258	0.1789	0.4331	294
16	0.0489	0.6163	224	0.1139	0.5076	272	0.1728	0.4374	310
17	0.0454	0.6178	237	0.1091	0.5101	287	0.1672	0.4413	326
18	0.0413	0.6214	249	0.1047	0.5124	302	0.1621	0.4449	342
19	0.0385	0.6226	262	0.0994	0.5163	316	0.1559	0.4496	357
20	0.0359	0.6236	275	0.0957	0.5181	331	0.1515	0.4525	373

shows the trade-off plot for considering A test data as 20% equivalent to B test data, with a prior distribution of $\pi \sim \text{beta}(29.8, 2.2)$. Now the PR is reduced to below 0.50, while the AP has dropped to below 0.65 for all constraint levels. The criteria values of the 21 optimal plans with different constraint levels on CR for using 40% and 20% historical information are summarized in Tables 2 and 3, respectively. In summary, a stronger connection between the historical data and the current test leads to a less diffuse prior distribution, which in turn leads to requiring fewer units to be tested in order to achieve the same level on the consumer's risk. Consequently, we can achieve a higher probability of passing the test with slightly more producer's risk. This is because the historical data have a higher success rate $144/150 = 0.96$ than the specified $\pi_0 = \pi_1 = 0.938$. In other words, the historical data support higher reliability, hence the more prior information is used, the fewer units need to be tested in the new demonstration test. On the other hand, if the historical data indicated lower reliability, then more units would need to be tested for a stronger prior distribution.

5. Discussion and Conclusions

Strategically choosing a best plan for conducting a demonstration test is of great practical importance in product design and innovation. The common practice of using the zero-failure test with minimum sample size for controlling only the consumer's risk can result in a test plan with unacceptably high risk for the producer and possibly a low probability of passing the test. Without good understanding of the implied levels of these other characteristics of the test, the choice of which test to use can be over-simplified and lead to an inferior choice that is not in the best interest of the different stakeholders.

In this paper, we have shown how enumeration of a large number of choices combined with quan-

titative evaluation to explicitly explore the interrelationships between four relevant criteria can lead to better understanding of choices when planning a demonstration test. We have focused on the consumer's and producer's risks, the acceptance probability for a successful test, and the cost. By examining a set of different choices of (n, c) values, useful conclusions are drawn regarding the general relationship between the design parameters and the test criteria, which can provide useful practical guidance for the practitioners facing similar problems. Quantifying the trade-offs between the consumer's and producer's risks when changing the sample size and/or the number of maximum allowable failures can help reshape the way the practitioners approach the general problems and encourage them to consider multiple aspects of their decision to make the best use of available resources.

Tactically, given the competing objectives for optimizing a demonstration test plan, we recommend a structured approach using a Pareto front to eliminate noncontending choices and to guide the process of making a quantitative and justifiable decision that match the goals of the test. Once a more manageable number of choices have been identified, then the practitioner can identify which of these choices most closely match their study goals. The approach presents strategies for making a decision based on different user priorities and practical/logistical constraints. For each strategy, the method sets a threshold for the primary criterion and finds the Pareto front based on simultaneously optimizing the remaining criteria. A set of graphical tools helps practitioners extract useful information and provides guidelines for making a further decision. It is worth noting that, among the three scenarios explored, controlling the consumer's risk first, which is the most common practice, can lead to a simple set of optimal solutions with a universal best plan to simultaneously optimize the remaining three criteria for each possible c value. Having found the set of superior solutions with the clear-cut trade-offs summarized in the trade-off plot in Figure 2 and Table 1, the user can make an easy and straightforward decision based on their requirements for PR and AP as well as how big a test they can afford. The other two scenarios involve richer trade-offs with a larger set of contending options. However, graphical tools provide compact summaries of useful information for supporting a tailored optimal decision. Despite that controlling consumer's risk primarily can lead to the smallest and simplest set of solutions, it is important for the practitioners to think carefully about what is most important for their particular application and choose the most appropriate approach to match their test goals. The R code for implementing the three decision-making strategies and generating the graphical summaries is available from the authors on request. Finally, the user-specified thresholds and prior distribution choices in a Bayesian analysis can have substantial impact on the final decision. We recommend explicitly exploring the subjective user inputs to gain a better understanding of their impact before making a final decision.

Note that, in demonstration test planning, there are many aspects that could be considered in the decision making. The producer's risk can be quantified by the probability of the producer rejecting a good product when it is actually good. This is useful to help a producer to make a direct decision on whether to release the product or not based on their level of tolerance for this risk. The cost of conducting the test is another important dimension in a decision, which is often a critical aspect to allow the producer to choose only an affordable test plan. However, a broader consideration of cost can also include the potential cost associated with a poor decision. For example, the cost from the producer's risk by rejecting a product can include the extra cost for re-inspecting the production process and redesigning and retesting the product, when it actually was sufficiently good. On the other hand, the cost from the consumer's risk by releasing an unacceptable product can include the extra cost generated due to product returns and the loss of customer loyalty, etc. A quantitative summary of these costs as consequences associated with producer's and consumer's risks can be evaluated based on their expectations over the posterior distribution of the reliability given different decisions. For example, the expected cost from the producer's risk (denoted by ECPR) can be quantified by

$$
\begin{aligned}
&\text{ECPR}\\
&\quad = E(\text{Cost} \mid \text{Test is failed})\\
&\quad = \left\{ \int_{\pi_0}^{1} \left[\sum_{y=c+1}^{n} \binom{n}{y} (1-\pi)^y \pi^{n-y} \right] p(\pi) C_1(\pi) d\pi \right\}\\
&\qquad \div \left\{ \int_{0}^{1} \left[\sum_{y=c+1}^{n} \binom{n}{y} (1-\pi)^y \pi^{n-y} \right] p(\pi) d\pi \right\},
\end{aligned}
$$

where $C_1(\pi)$ is the cost associated with the producer's risk when reliability is at π for $\pi \geq \pi_0$. Similarly, the expected cost from the consumer's risk (de-

noted by ECCR) can be calculated as

$$\begin{aligned}
&\text{ECCR}\\
&\quad = E(\text{Cost} \mid \text{Test is passed})\\
&\quad = \left\{ \int_0^{\pi_1} \left[\sum_{y=0}^{c} \binom{n}{y} (1-\pi)^y \pi^{n-y} \right] p(\pi) C_2(\pi) d\pi \right\}\\
&\qquad \div \left\{ \int_0^{1} \left[\sum_{y=0}^{c} \binom{n}{y} (1-\pi)^y \pi^{n-y} \right] p(\pi) d\pi \right\},
\end{aligned}$$

where $C_2(\pi)$ is the cost associated with the consumer's risk when reliability is at π for $\pi \le \pi_1$. These expected costs can be approximately estimated based on Monte Carlo integration using

$$\begin{aligned}
&\widehat{\text{ECPR}}\\
&\quad = \left\{ \sum_{j=1}^{M} \left[1 - \sum_{y=0}^{c} \binom{n}{y} (1-\pi^{(j)})^y (\pi^{(j)})^{n-y} \right] \right.\\
&\qquad\qquad \left. \times I(\pi^{(j)} \ge \pi_0) C_1(\pi^{(j)}) \right\}\\
&\qquad \div \left\{ \sum_{j=1}^{M} \left[1 - \sum_{y=0}^{c} \binom{n}{y} (1-\pi^{(j)})^y (\pi^{(j)})^{n-y} \right] \right\}
\end{aligned}$$

and

$$\begin{aligned}
&\widehat{\text{ECCR}}\\
&\quad = \left\{ \sum_{j=1}^{M} \left[\sum_{y=0}^{c} \binom{n}{y} (1-\pi^{(j)})^y (\pi^{(j)})^{n-y} \right] \right.\\
&\qquad\qquad \left. \times I(\pi^{(j)} \le \pi_0) C_2(\pi^{(j)}) \right\}\\
&\qquad \div \left\{ \sum_{j=1}^{M} \left[\sum_{y=0}^{c} \binom{n}{y} (1-\pi^{(j)})^y (\pi^{(j)})^{n-y} \right] \right\}.
\end{aligned}$$

Then, the estimated expected cost from the producer's and consumer's risks can be used separately or in combination to give an estimated overall expected cost for selected test plans.

References

Anderson-Cook, C. M. and Lu, L. (2015). "Much-Needed Structure: A New 5-Step Decision-Making Process Helps You Evaluate, Balance Competing Objectives". *Quality Progress* 48(10), pp. 42–50.

Deb, K. (2009). *Multi-Objective Optimization Using Evolutionary Algorithms*. New York, NY: Wiley.

Easterling, R. G. (1970). "On the Use of Prior Distributions in Acceptance Sampling". *Annals of Reliability and Maintainability* 9, pp. 31–35.

Gelman, A.; Carlin, J. B.; Stern, H. S.; and Rubin, D. B. (2004). *Bayesian Data Analysis*, 2nd edition. Boca Raton, FL: Chapman and Hall/CRC.

Gronwald, W.; Hohm, T.; and Hofmann, D. (2008). "Evolutionary Pareto-Optimization of Stably Folding Peptides". *BMC-Bioinformatics* 9, p. 109.

Hamada, M. S.; Wilson, A. G.; Reese, C. S.; and Martz, H. F. (2008). *Bayesian Reliability*. Springer.

Hart, L. (1990). "Reliability of Modified Designs: A Bayes' Analysis of an Accelerated Test of Electronic Assemblies". *IEEE Transactions on Reliability* 39, pp. 140–144.

Kasprzak, E. M. and Lewis, K. E. (2001). "Pareto Analysis in Multiple Optimization Using the Collinearity Theorem and Scaling Method". *Structural Multidisciplinary Optimization* 22, pp. 208–218.

Lange, K. (2013). *Optimization*, 2nd edition. New York, NY: Springer.

Lu, L.; Anderson-Cook, C. M.; and Robinson, T. J. (2011). "Optimization of Designed Experiments Based on Multiple Criteria Utilizing a Pareto Frontier". *Technometrics* 53, pp. 353–365.

O'Connor, P. and Kleyner, A. (2012). *Practical Reliability Engineering*, 5th edition. Cambridge, UK: Wiley-Blackwell.

Pintar, A.; Lu, L.; Anderson-Cook, C. M.; and Silver, G. L. (2012). "Bayesian Estimation of Reliability for Batches of High Reliability Single-Use Parts". *Quality Engineering* 24 (4), pp. 473–485.

Trautmann, H. and Mehnen, J. (2009). "Preference-Based Pareto Optimization in Certain and Noisy Environments". *Engineering Optimization* 41, pp. 23–38.